Transportation Research Interdisciplinary Perspectives 6 (2020) 100176

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: https://www.journals.elsevier.com/transportation-research-interdisciplinary-perspectives

Vehicle involvements in hydroplaning crashes: Applying interpretable machine learning

Subasish Das [a,*], Anandi Dutta [b], Kakan Dey [c], Mohammad Jalayer [d], Abhisek Mudgal [e]

[a] Texas A&M Transportation Institute, 1111 RELLIS Parkway, Room 4414, Bryan, TX 77807, United States of America
[b] Department of Computer Science, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249-0667, United States of America
[c] Civil and Environmental Engineering, West Virginia University, 1374 Evansdale Drive, Morgantown, WV 26506, United States of America
[d] Department of Civil and Environmental Engineering, Rowan University, Glassboro, NJ 08028, United States of America
[e] Department of Civil Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh 221005, India

ARTICLE INFO

Article history:
Received 3 March 2020
Received in revised form 21 May 2020
Accepted 10 July 2020
Available online 23 July 2020

Keywords:
Hydroplaning crash
Crash narrative
Machine learning
Precision
Recall

ABSTRACT

Although hydroplaning is a major contributor to roadway crashes, it is not typically reported in conventional crash databases. Hence, a framework to classify various crash attributes from police reports and to identify hydroplaning crashes is strongly needed. This study applied natural language processing (NLP) tools to seven years (2010–2016) of crash data from the Louisiana traffic crash database to identify hydroplaning related crashes. This research focused on the development of a framework to apply interpretable machine learning models to unstructured textual content in order to classify the number of vehicle involvements in a crash. This approach evaluated the effectiveness of keywords in determining the classification. This study used three machine learning algorithms. Of these algorithms, the eXtreme Gradient Boosting (XGBoost) model was found to be the most effective classifier. This research provided a platform to understand the application of interpretability in machine learning models. The outcomes of this study prove that underlying trends or precursors can be revealed and analyzed through these models. Furthermore, this indicates that quantitative modeling techniques can be used to address safety concerns.

1. Introduction

Hydroplaning is a dangerous phenomenon that can cause roadway crashes and result in severe injuries or fatalities. The term hydroplaning is often used by roadway users to describe a wide variety of wet-roadway driving hazards. However, the more accurate definition of hydroplaning describes a specific condition of tire skidding on wet pavement in which the automobile tires float on a layer of water at a high speed, causing the driver to lose control of the vehicle. This occurs when the water underneath the tire cannot escape, and hydrodynamic pressure builds up to sustain the floating tire over the area. Some common contributing factors of hydroplane crashes include vehicle speed, tire pressure, tread depth, pavement texture, and roadway and environmental conditions (Federal Highway Administration, n.d.).

Previous research on hydroplaning crashes has followed conventional methods to establish relationships between crash frequency and key contributing factors, such as the ones mentioned above (Enustun, 1976; Black and Jackson, 2000; Aycock, 2008; Gunaratne et al., 2012; Jayasooriya and Gunaratne, 2014; Zhou et al., 2019). Many recent studies have focused on determining which factors significantly affect different crash characteristics. Conventional data analysis methods perform crash frequency or injury severity analysis on police-reported crash data. However, these reports usually contain a textual description of the crash event that has not been explored. Information from these narratives can require significant effort to understand the event circumstances as they are not stored electronically, and they are presented as unstructured or semi-structured free-text data format. Because of this, precision information can be lost through manual and time-consuming interpretation of these text reports.

Text mining has become an increasingly popular research area since it is effective in identifying valuable patterns and hidden insights from plain texts. This study used text mining and machine learning (ML) algorithms to investigate hydroplaning crashes and develop text mining-based models to determine key crash contributing factors such as road, vehicle, and environmental conditions. This study aims to evaluate various text mining classification techniques by measuring their ability to classify crash narratives for seven years of crash data obtained from the Louisiana traffic crash database. The study evaluated three ML algorithms: support vector machine (SVM), random forest (RF), and eXtreme Gradient Boosting (XGBoost).

The structure of this paper is as follows: the literature review provides a comprehensive overview of previous studies focused on hydroplane crashes. The paper then introduces the modeling tools used in the study.

* Corresponding author.
E-mail addresses: s-das@tti.tamu.edu, (S. Das), anandi.dutta@utsa.edu, (A. Dutta), kakan.dey@mail.wvu.edu, (K. Dey), abhisek.civ@iitbhu.ac.in. (A. Mudgal).

http://dx.doi.org/10.1016/j.trip.2020.100176

The next sections demonstrate the data preparation and application of text mining techniques. Finally, this paper describes the results of this evaluation, followed by the conclusions and discussion of the study.

2. Literature review

Most of the studies included in this review focus on theoretical concept development for hydroplaning occurrences (Enustun, 1976; Black and Jackson, 2000; Aycock, 2008; Gunaratne et al., 2012). There are a limited number of studies that aim to determine the key contributing factors of hydroplaning related crashes.

2.1. Studies on hydroplaning crashes

In 1976, Enustun (1976) conducted a study that aimed to avoid the recurrence of identified hydroplaning crashes. To achieve this, Enustun identified the locations of crashes associated with hydroplaning to examine the road, vehicle, and weather conditions related to hydroplaning. The study concluded that true hydroplaning crashes on highways were rare and made up less than 5% of wet-pavement crashes. Enustun suggested that a greater effort be made to educate the public about the degree of tire capability deterioration on wet pavements under rapid speeds. Black and Jackson (2000) conducted a study examining the phenomenon of hydrodynamic drag, which is virtually anonymous to highway and transportation engineers. Aycock (2008) investigated a 2003 hydroplaning crash on an interstate highway in Georgia and concluded that the speed of the vehicle, the condition of the tires, and the depth of the water on the roadway all contributed to the hydroplaning crash. Gunaratne et al. (2012) developed statistical models that estimated wet weather speed reduction as well as analytical and empirical methods to predict hydroplaning speeds of trailers and heavy trucks. Using crash data, geometrical data, pavement condition data, and other relevant information available for Florida roadways, the study conducted a wet weather crash analysis. The results of this study indicated that wider sections increased the likelihood of hydroplaning crashes, and dense-graded pavements were more likely to induce conditions fit for hydroplaning than open-graded ones. In a follow-up study, Jayasooriya and Gunaratne (2014) developed statistical models to estimate the thickness of water film that develops on roadways and classified the corresponding threshold hydroplaning speed into empirical, analytical, and numerical categories. Zhou et al. (2019) conducted a study to measure the impacts of automated vehicles (AV) on roadway hydroplaning and pavement life in comparison to human-driven vehicles. The results showed that a uniformly distributed lateral wandering pattern for AVs prolonged pavement fatigue life, decreased pavement rut depths, and reduced hydroplaning potential.

2.2. Text mining studies on crash narratives

Text mining, an extremely useful data analysis technique, has been used to extract valuable information from large text-based datasets. Text mining methods can identify patterns and anomalies in data over time, determine key contributing factors, and develop predictive models to solve real-world problems (Gopalakrishnan and Khaitan, 2017; Jin et al., 2007). In-depth text mining has primarily been used in crash and injury analysis to gain insights from occupational crash reports (Abdat et al., 2014; Marucci-Wellman et al., 2011; Smith et al., 2006; Bertke et al., 2016; Vallmuur et al., 2016; Bondy et al., 2005; Bunn et al., 2008; Williamson et al., 2001), health care reports (Chen et al., 2016; Marucci-Wellman et al., 2015; Wang et al., 2012), automobile crash reports (Marucci-Wellman et al., 2011; Smith et al., 2006; Bertke et al., 2016; Vallmuur et al., 2016; Bondy et al., 2005; Bunn et al., 2008; Williamson et al., 2001; Chen et al., 2016; Marucci-Wellman et al., 2015; Wang et al., 2012), and others (Gopalakrishnan and Khaitan, 2017; Jin et al., 2007; Brown, 2016).

Concept Chain Queries, a special case of text mining, is used to identify essential evidence trails across documents to explain relationships between two topics of interest (Jin et al., 2007). Researchers have developed a text search technique to explore the utility of crash narrative text analysis for generating code to determine injury mechanisms (Williamson et al., 2001). Haddon's matrix provided a conceptual framework for researchers to code text from work-related injury reports to identify the contributing factors of these injuries (Bondy et al., 2005). In another study, Haddon's matrix separated the fatal incident into three event phases (pre-event, event, post-event) and provided a coded data set as well as coding rules (Bunn et al., 2008).

Although the crash data are small and fairly homogenous, it is challenging to recognize patterns within the narrative text. A combined Naïve-Fuzzy Bayesian approach can be used to provide a more accurate narrative classification and to identify the data that require manual review, thus reducing the burden on human coders (Abdat et al., 2014; Marucci-Wellman et al., 2015; Graves et al., 2015). Researchers have also utilized DUALIST to easily classify data; DUALIST is an online interactive program that allows novice users to quickly classify thousands of narratives after approximately 60 min of training (Gopalakrishnan and Khaitan, 2017). Studies have also evaluated the effectiveness of the Bayesian-based model in comparison to other ML algorithms. The Bayesian-based model was compared to the neural network (Gopalakrishnan and Khaitan, 2017), logistic regression model (Pollack et al., 2013; Sorock et al., 1996), and support vector machine model (Sorock et al., 1996), all of which provided higher accuracy in classifying the emerging causes of occupational injury.

Another study compared a Semi-Supervised Set Covering Machine (S3CM) learning algorithm to the Transudative Vector Support Machine (TVSM), the original fully supervised Set Covering Machine (SCM), and 'Freetext Matching Algorithm' natural language processor. The S3CM algorithm was developed to detect the presence of coronary angiograms and ovarian cancer diagnoses from electronic health record narratives. The model does not rely on linguistic rules and worked effectively after being trained with pre-classified test data sets. The study found that S3CM results were better than TVSM and fully supervised SCM (Williams and Betak, 2016).

Text mining has become an increasingly popular crash analysis method in the area of transportation engineering. A study demonstrated a connectionist-based model to classify free-text crash descriptions (Chatterjee, 1998); singular value decomposition was used for feature extraction and network training. This model was evaluated in comparison to a fuzzy Bayes model and a keyword-based model. All three models were applied to human classified data, and the study found that the connectionist and fuzzy Bayes model outperformed the keyword model. Another study performed exploratory text mining and empirical Bayes (EB) data mining to understand the associations between vehicle condition and automotive safety (Das et al., 2018). These methods were used to identify key crash causing factors in terms of vehicular manufacturing defects (e.g., airbags, brake system, seat belts, and speed control).

Researchers developed a three-level hierarchical Bayesian model, named Latent Dirichlet Allocation (LDA), to identify major recurring crash topics from textual data compiled from the Federal Railroad Administration (FRA) reports. The researchers then applied a text clustering method to the text using Jigsaw text visualization software and found that both methods had equal effects. In another study, LDA, RF, and partial least squares techniques were combined and applied to identify the contributing factors and to accurately predict the cost of railroad crashes (Pollack et al., 2013). Researchers also used logistic regression (Fitzpatrick et al., 2017; Graves et al., 2015; Williams and Betak, 2016), clustering (Williams and Betak, 2016), and other computational tools (e.g., Statistical Analysis System (SAS) and Leximancer) (Nayak et al., 2010; Sorock et al., 1996) to identify vehicle crash factors. In summary, computerized approaches and predictive models can be used to standardize crash narrative text analysis and reduce human error in crash and injury surveillance.

The literature review indicates that there is a need for further research and an in-depth investigation on hydroplaning crashes. As hydroplaning crashes can often involve multiple-vehicles, this study aims to identify the trends from crash narratives by performing ML and classifying the nature of single and multiple-vehicle involvement in hydroplaning crashes.

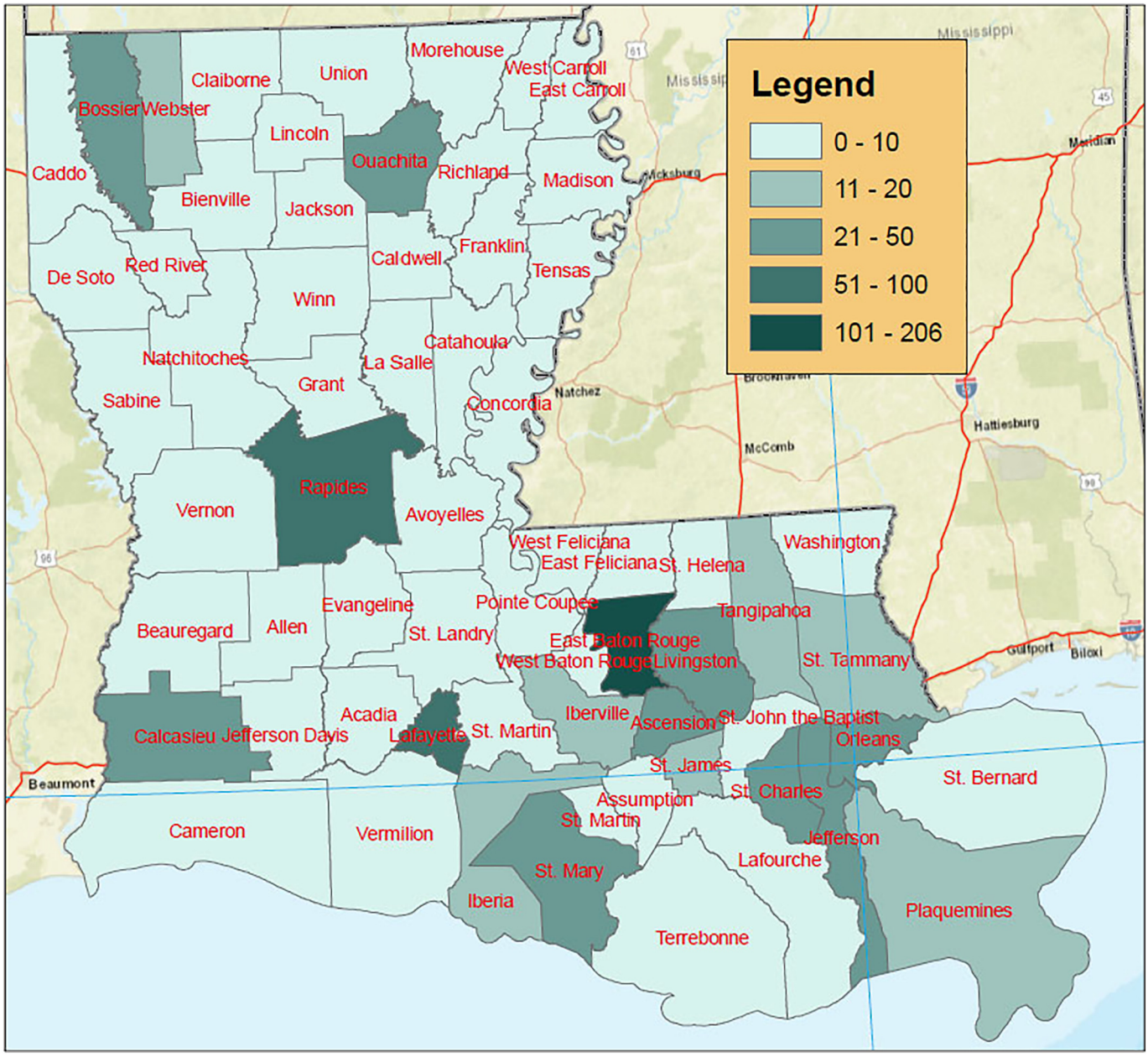


Fig. 1. Hydroplaning crashes by Parishes.

3. Methodology

3.1. Defining hydroplaning crashes

Hydroplaning refers to uncontrolled sliding of a vehicle on the wet road surface. It is triggered when a tire experiences more water than it can dissipate between the tire tread and highway surface bed, resulting in traction failure. The pressure of water generated in the front wheel forces the water to wedge under the tire's foremost edge causing the tire to get lifted from the highway surface. The result makes the tires skate on the water sheet with little or no contact on the highway surface, which makes the driver lose control of vehicle. The result is a hydroplaning crash (*Traction friction of tires by Ron Kurtus - physics lessons: school for champions*, n.d.; *Roadway hydroplaning - the trouble with highway cross slope*, n.d.). Yeager (1974) examined tire hydroplaning and its effect on wet traction and gathered a wide range of data.

3.2. Machine learning models

ML models are gaining popularity among the researchers due to the high prediction power and flexibility. A ML model is created by a ML algorithm or a set of rules that helps in learning how to predict a specific outcome based on past events. ML can identify patterns in the data and make predictions based on a training process. Once a model is developed, it can be used to predict measures or classify types based on the research question. In comparison to ML, conventional statistical modeling uses mathematical equations to identify associations between variables. It is also significantly beneficial because it has a higher interpretability than ML models. One major disadvantage of statistical modeling is its dependence on predetermined assumptions. The major disadvantage of ML models is that it is more challenging to comprehend or describe the outcomes (Friedman, 2001; Molnar, 2018). Natural Language Processing (NLP) and Text Mining (TM), in conjunction with ML, are an efficient data-centric tool to collect,

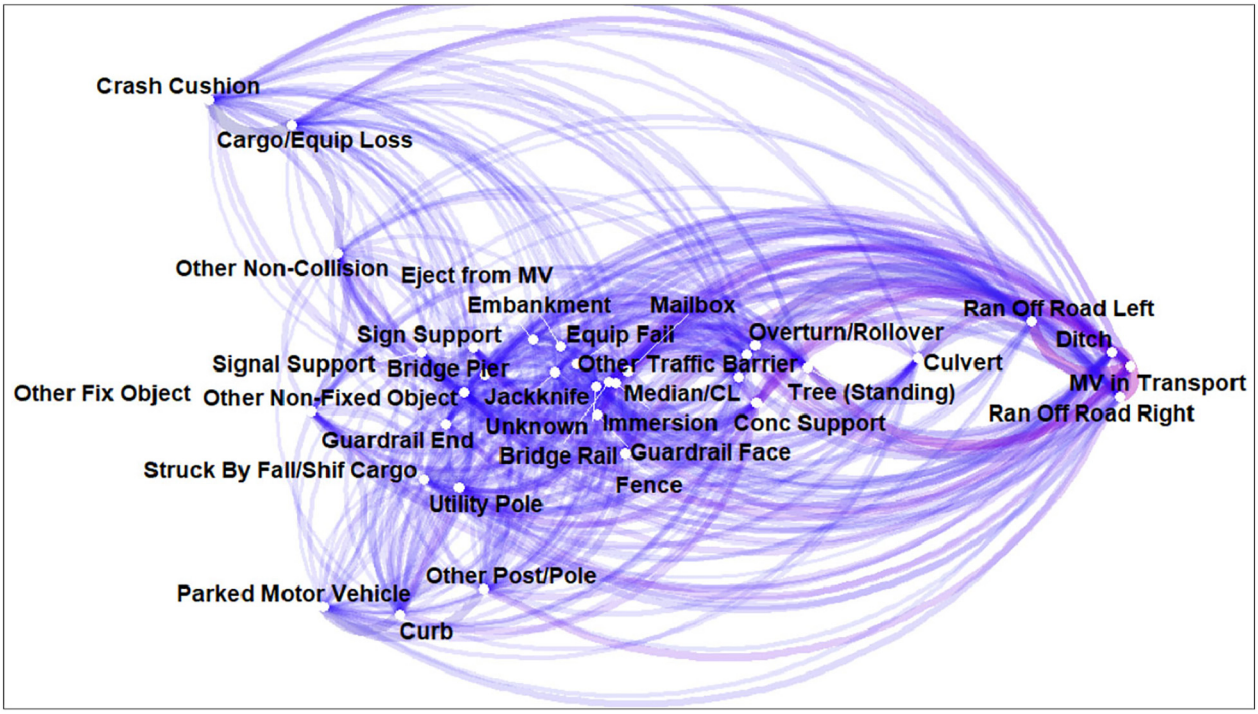


Fig. 2. Association between sequences of harmful events in hydroplaning crashes.

analyze, and extract interesting findings or patterns from big data such as the collection of crash narrative reports. This study performed multiple longitudinal studies to illustrate various interesting patterns and anomalies in the data using text mining pipelines. The classification experiment designed in this study included three ML classifiers: RF, SVM, and XGBoost. Interested readers can consult Bishop's book for a more complete review of these algorithms (Bishop, 2011).

3.2.1. Random forest (RF)

The framework of the random forest (RF) algorithm is based on the bagging principle (Breiman, 2001) and random subspace method (Ho, 1998). The core step involves developing a collection of decision trees with random predictors. The two important byproducts in this process are 1) out-of-bag error rate (OOB) and 2) variable importance measures (VIM). The OOB is known as the misclassification rate, and it decreases as the number of trees increase. The number of trees does not affect overfitting the data; therefore, a significant number of trees can be used. To decrease the bias and correlation, the trees are grown to their maximum depth (Ho, 1998). VIM can be measured by the classification accuracy and Gini impurity (the criterion that each child node reaches its highest purity, with all observations on that child node belonging to the same classification).

3.2.2. Support vector machine (SVM)

In 1963, Vapnik and Lerner introduced the concept of Generalized Portrait algorithm (*History of SVM*, n.d.). It has been considered as the core conceptual framework for developing support vector machine (SVM). Later, Vapnik started the field of statistical learning theory in 1974 (*History of SVM*, n.d.). Vapnik et al. introduced the current form of the SVM on a basis of a separable bipartition problem at the AT & T Bell Laboratories in 1992 (Smola and Schölkopf, 2004). It adopts the theory of minimizing structural risk instead of minimizing empirical risk (also known as training error). SVM can be applied for solving both classification and regression problems. By expressing the regression or classification outcomes in terms of linear combination (using training data), it generates a fraction of data points known as support vectors (with non-zero co-efficient). The key concept of SVM is to map the data x into a high-dimensional feature space F via a nonlinear mapping (Cornejo-Bueno et al., 2016).

3.2.3. Extreme Gradient Boosting (XGBoost)

Tree boosting is a very effective ML tool. XGBoost implements ML algorithms under the framework of gradient boosting (Friedman, 2001). It provides a parallel tree boosting algorithm that reaches to the optimized point in a fast and accurate way. This framework has several advantages such as efficient regularization parameters, early stopping, parallel processing, effective loss functions, and different base learners. The success of XGBoost relies on its scalability.

3.3. Interpretability

Statistical models are widely popular based on its interpretability. Modeling framework with adequate interpretability can help users comprehend how the estimations were determined (Lundberg and Lee, 2016). The supervised ML algorithms are trained to predict the outcome variable based on the training data. Due to its keen focus on improving precision, the interpretability part has been usually not prioritized. For example, an algorithm aiming to determine a non-motorist in the traveling path of an autonomous vehicle solely focuses on high prediction accuracy. Interpretation might not be the key focus in such case as inaccuracy of the algorithm can involve a fatal or serious injury. However, interpretation is significantly important for some cases. A model with high interpretability means that the model can explain its decision, and the end-users can understand how to improve the decision-making function. In many cases, interpretability can be used as a decision making tool. In summary, an interpretable model can ensure fairness, data protection, consistency, interconnection, and trust (Fisher et al., 2018; Molnar, 2018).

4. Data description

4.1. Data preparation

The dataset of the current study is comprised of crash narrative reports in text format that were collected from police-reported crashes in Louisiana

from 2010 to 2016. The current dataset does not provide filtering options for hydroplaning related crashes. The researchers used a text searching algorithm to identify crashes associated with the keywords: 'hydroplane' and 'hydroplaning.' Additionally, word stemming was conducted to make the search robust. Initially, the search algorithm identified 703 crash reports that might be related to hydroplaning. Undergraduate students inspected those crash reports and removed reports that were out of the scope of hydroplaning related crashes. This resulted in a total of 652 remaining crash reports.

Fig. 1 illustrates the spatial distribution of hydroplaning crashes for different parishes. Out of 64 Parishes, 51 Parishes experience at least one hydroplaning crash during 2010–2016. The illustration also shows that East Baton Rouge has a high number of hydroplaning crashes.

4.2. Contexts of single-vehicle and multiple-vehicle hydroplaning crashes

One of the most important aspects of understanding hydroplaning crashes is to comprehend the mechanism of crash events. It is also important to know the patterns of single and multiple-vehicles crashes associated with hydroplaning. Louisiana crash data provides sequences of harmful events for each crash records. These sequences are:

- First harmful event
- Second harmful event
- Third harmful event
- Fourth harmful event.

Fig. 2 illustrates the network plot of the association between the sequences of the harmful events for the hydroplaning crashes. The darkness of the lines indicates the node density associated with the harmful event types. The nodes on the right (motor-vehicle or MV in transport or multiple-vehicle crash, ran off-road, and ditch) show higher density. Most of the harmful events are associated with single-vehicle crashes or ran off-road crashes; however, multiple-vehicle crashes (struck by MV in transport) are also common. The final dataset shows that 32% of hydroplaning crashes are multiple-vehicle crashes. A study done by Khattak et al. also showed similar statistics (Khattak et al., 1998).

Table 1 lists the chi-square test values (a convenient test to determine the difference between the dataset attributes) and descriptive statistics of the key variables (variables listed here have p-value < 0.05). The p-values from the chi-square tests indicate that some of the major contributing factors are significantly different for two types of crashes (single vehicle vs. multiple-vehicle) that are associated with hydroplaning. Multiple-vehicle hydroplaning crashes are higher in percentage than single vehicle crashes. Interstate crashes are high in single vehicle hydroplaning crashes. Multiple-vehicle hydroplaning crashes are high in percentages in normal condition (roadways with no abnormalities, normal driving, and non-distracted drivers). The next section shows how interpretable machine learning (IML) can be applied in classifying single and multiple-vehicle hydroplaning crashes.

As this study is focused on the identification of single or multiple-vehicle collisions, it is important to examine the significant vehicle related variables. Vehicle type, vehicle condition, and vehicle year are considered as the key variables that can provide a glimpse of the vehicle condition and its impact on hydroplaning related crashes. Vehicle types are not significantly different in single and multiple crash groups. However, passenger car crashes are higher in single vehicle crashes. Engine failure is the key factor of the vehicle condition types. For multiple-vehicle crashes, this percentage is higher than single vehicle crashes. Single vehicle crashes show slightly higher percentages than multiple-vehicle crashes. The vehicle year variable also does not show significant differences among single and multiple-vehicle crashes. Percentages of both new (2011 and after) and old model (2000 and before) cars are disproportionately higher in multiple crash groups (Table 2).

Table 1
Chi squared tests and descriptive statistics for key variables.

Attributes	Multiple	Single	p-Value
	N = 575	N = 566	
Access control			< 0.001
No control (unlimited access to roadway)	70.70%	58.70%	
Full control (only ramp entrance & exit)	20.20%	33.20%	
Partial control (limited access to roadway)	9.06%	7.60%	
Other	0.00%	0.35%	
Unknown	0.00%	0.18%	
Highway type			< 0.001
Interstate	28.90%	38.70%	
State hwy	24.00%	22.00%	
U.S. hwy	16.40%	19.30%	
Parish road	11.10%	12.80%	
City street	19.50%	7.27%	
Road condition			< 0.001
No abnormalities	78.60%	60.20%	
Water on roadway	20.70%	36.70%	
Bumps	0.35%	0.71%	
Deep ruts	0.00%	0.35%	
Other	0.02%	0.54%	
Weather			< 0.001
Rain	78.90%	90.80%	
Cloudy	12.90%	6.54%	
Clear	7.84%	1.41%	
Fog/smoke	0.35%	0.35%	
Other	0.00%	0.53%	
Sleet/hail	0.00%	0.35%	
Day of the week			0.023
FSS	43.00%	49.80%	
MTWT	57.00%	50.20%	
Driver condition			< 0.001
Normal	71.30%	61.20%	
Inattentive	23.80%	32.40%	
Apparently asleep/blackout	0.28%	0.46%	
Unknown	2.98%	3.19%	
Driver distraction			0.001
Not distracted	82.00%	75.40%	
Cell phone	0.35%	0.18%	
Other electronic device (pager, palm pilot, navigation device, Etc.)	0.01%	0.18%	
Other inside the vehicle	0.01%	0.71%	
Other outside the vehicle	2.29%	1.06%	
Unknown	15.30%	22.50%	
Driver gender			0.001
F	41.00%	36.70%	
M	55.80%	62.70%	
U	3.13%	0.53%	

Table 2
Key vehicle related variables.

Variable	Multi	Single
Vehicle type		
Passenger car	46.38%	53.39%
Lt. truck (P.U., etc.)	27.15%	29.19%
Suv	18.78%	15.84%
Other	7.69%	1.81%
Vehicle condition		
Engine failure	92.99%	88.24%
Defective brakes	0.45%	0.23%
Worn or smooth tires	1.58%	4.07%
Unknown	4.98%	7.69%
Vehicle years		
2000 and before	16.06%	13.35%
2001–2010	57.47%	64.71%
2011 and after	26.47%	22.17%

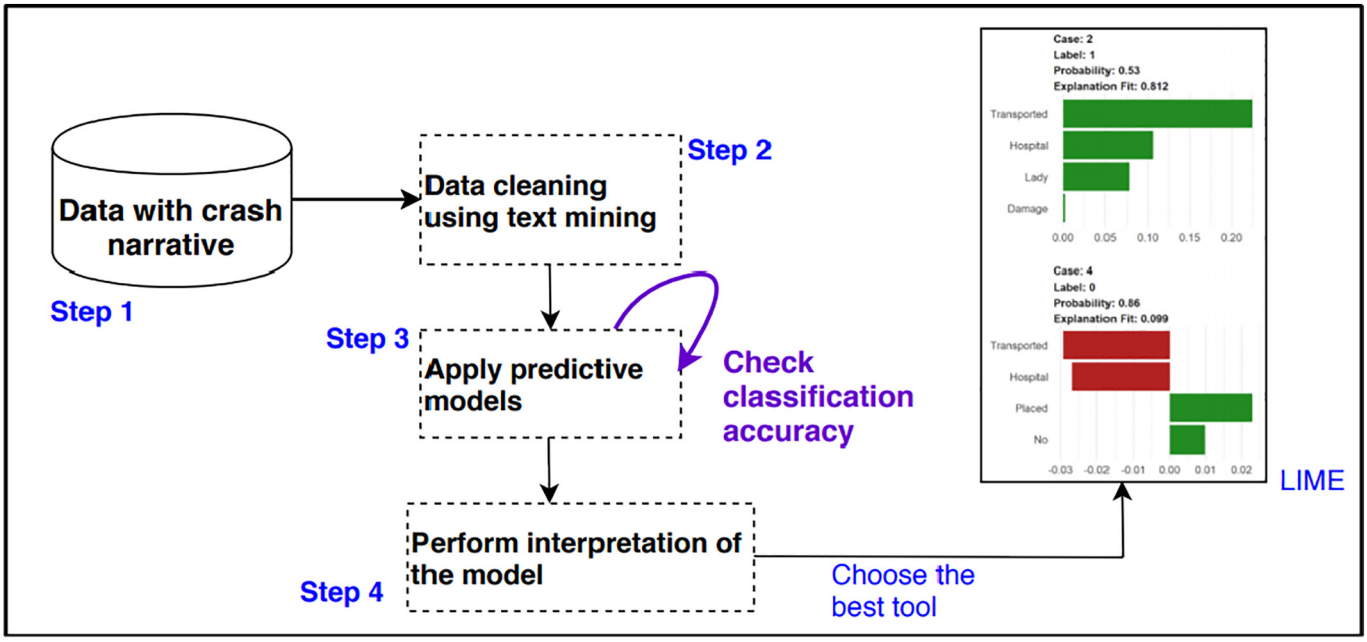


Fig. 3. Framework for crash narrative analysis.

4.3. Framework for crash narrative analysis

This study developed a framework, as shown in Fig. 3, for the application of IML in solving the classification problem using crash narrative data. The steps are the following:

- **Step 1: Data collection.** This study collected the electronic data of the crash narratives. In many cases, the crash reports are hand-written and are not electronically recorded. Louisiana has begun the process of transferring all crash reports to electronic versions. The current dataset has crash narratives for at least 50% of the crash records starting from 2010.
- **Step 2: Data cleaning.** Data cleaning can be performed by using text mining algorithms. Available lexicons were used to remove redundant words, but there is still a need for domain-specific lexicons. For example, vehicle numbers (i.e., vehicle 1, vehicle 2) are associated with the determination of the involvements of vehicles. A general lemmatization (i.e., removal of inflectional endings from words) was performed. However, some of the significant key words were not lemmatized due to the importance of meaning in different scenarios. For example, 'stop' may indicate a 'stop sign,' but 'stopped' may be associated with two-vehicle crashes due to the sudden stoppage of the front vehicle.
- **Step 3: Application of IML model.** This study applied three different ML models to determine the most suitable model. Many IML tools have been developed in the recent years. Some of the popular methods are partial dependence plot (PDP), Local interpretable model-agnostic explanations (LIME), feature interaction, feature importance, global surrogate models, Individual conditional expectation (ICE), and Shapley value explanations (Friedman, 2001; Friedman and Popescu, 2018; Fisher et al., 2018; Ribeiro et al., 2016; Lundberg and Lee, 2016; Pedersen and Benesty, n.d.). In this study, this study used LIME as an IML tool.

Before applying the ML algorithms, the text mining tools were applied to make the dataset less noisy. The most common problem in text data analysis is the excessive number of terms and redundant features in a single narrative. Additionally, words or parts of the words with similar meaning are compressed into the same term (known as word stemming) to make the classification more robust. This study performed a basic redundant word (i.e., removal of prepositions, articles, common words such as road, crash) to prepare the final dataset. Future work may include more robust data cleaning that improves model precision based on domain-specific lexicons and stop word list. Fig. 4 shows a sample example of the original versus the modified crash narrative.

5. Modeling results

Three databases were developed for the analysis: 1) a train set with 452 crashes (67%), 2) a validation set with 88 crashes (13%), and 3) a test set with 112 crash records (17%) using stratified resampling. After several trial and errors, this study used several thresholds for the models. For example, the used parameters in XGBoost model are: 1) step size shrinkage = 0.10, 2) minimum loss reduction = 0.01, 3) maximum depth of a tree = 10, and 4) regularization weights = 0.2, and 1. Models developed from train data were applied to validation and testing data to evaluate their performance. Table 3 lists the model outcomes in classifying single and multiple-vehicles hydroplaning crashes. The values show that XGBoost algorithm performed better than both SVM and RF. Accuracies were calculated as 84%, 77%, and 71% for train, validation, and test data, respectively. These accuracies can be further improved by developing a robust crash narrative lexicon that includes stop words and trigger words, or words that are highly associated with crash outcomes.

True positive (TP) and false positive (FP) are the measures of correct and incorrect classifications per actual or real class, respectively. True negative (TN) and false negative (FN) are measures of correct and incorrect rejections per real class, respectively (Labatut and Cherifi, 2011). Some of the common measures of model performance and accuracy are listed in Table 4.

Table 5 lists the performance measures of the ML models for train, validation, and test data. Based on the performance measures described above, XGBoost shows better performances than the other two models.

As previously mentioned, "trusting a prediction" is important to "trust the model." However, the confusion matrix may not always be suitable to evaluate the model; hence, model explanation is crucial. This study applied a recently developed interpretable ML algorithm, LIME, to explain the predictions of any classifiers in an interpretable manner. This study used the R package 'lime' to interpret the model outcomes (Pedersen and Benesty, n.d.). For example, considering the target 'multiple-vehicle crash,' six crash reports are randomly selected to visualize the association of the keywords and the output (Fig. 5a). Label one will indicate that the XGBoost model predicts the crash as a multiple-vehicle hydroplaning crash. Similarly, label zero will indicate the crash as a single-vehicle crash. The explanation models

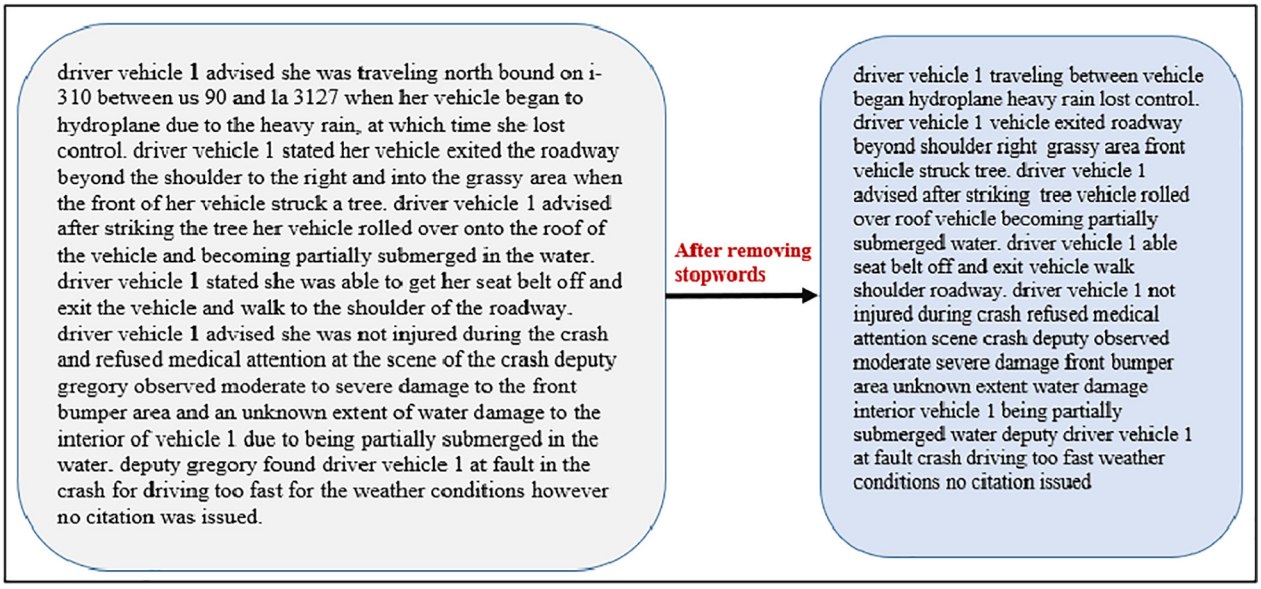


Fig. 4. Sample example of original and cleaned text narrative.

considered the 10 words with the highest associations in this study (blue indicates the word supports in true classification and red indicates the word contradicts in the classification task). For explanation plots, only the top four highly associated words were displayed. Fig. 5a shows that all cases (except Case 3) are single-vehicle crashes; the bars indicate the association with the classification. For example, Case 1 shows that the probability is 0.94

Table 3
Confusion matrix of the predicted classes.

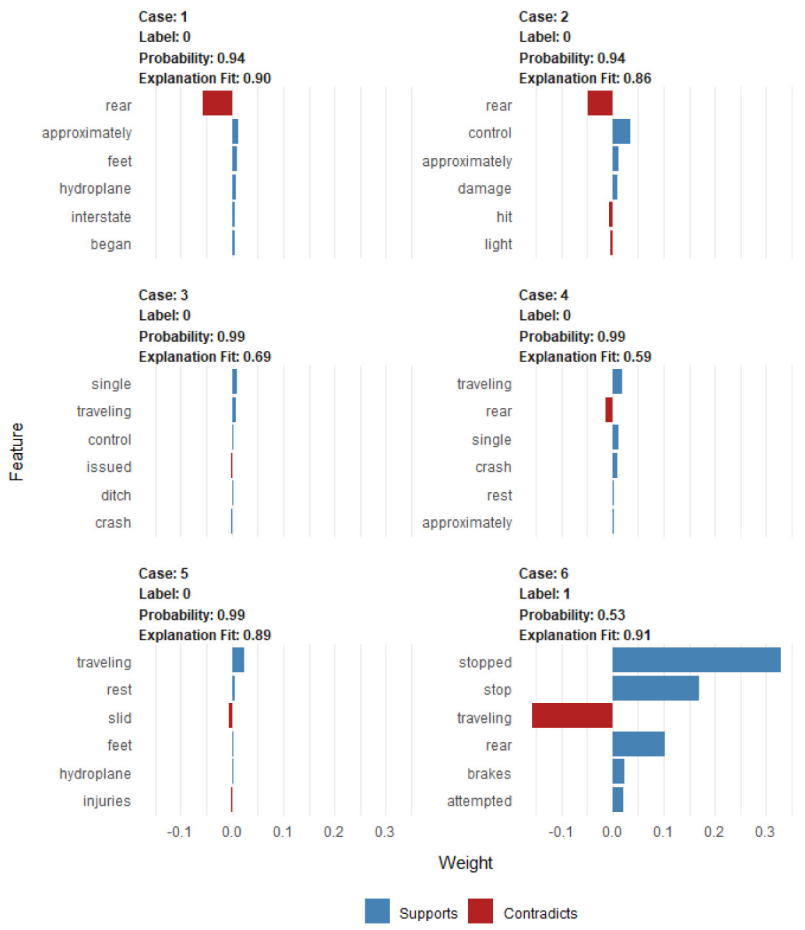
Model	Class (observed)	Train (452 crashes)		Validation (88 crashes)		Test (112 crashes)	
		Single (predicted)	Multiple (predicted)	Single (predicted)	Multiple (predicted)	Single (predicted)	Multiple (predicted)
SVM	Single	255	57	34	20	46	31
	Multiple	39	101	11	23	12	23
RF	Single	260	52	36	18	48	29
	Multiple	38	102	9	25	11	24
XGBoost	Single	270	42	40	14	54	23
	Multiple	30	110	6	28	9	26

Table 4
Performance measures.

Measure	Definition	What is measured
Recall or sensitivity	$\frac{TP}{TP+FN}$	Effectiveness of positive level identifications
Specificity	$\frac{TN}{TN+FP}$	Effectiveness negative level identifications
Precision	$\frac{TP+TN}{TP+TN+FP+FN}$	Class agreement of the data labels with the positive labels
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall accuracy
Balanced Accuracy	$\frac{TP}{TP+FN} \times 0.5 + \frac{TN}{TN+FP} \times 0.5$	Average of the proportion corrects of each class individually
$F-score$	$\frac{2 \times Precision \times Recall}{Precision+Recall}$	Weighted average of the recall and precision

Table 5
Performance measures if the ML models.

Model	Data	Sensitivity	Specificity	Accuracy	Balanced accuracy	Precision	F-score
SVM	Train	0.8173	0.7214	0.7876	0.7694	0.8673	0.8416
	Validation	0.6296	0.6765	0.6477	0.6531	0.7556	0.6869
	Test	0.5974	0.6571	0.6161	0.6273	0.7931	0.6815
RF	Train	0.8333	0.7286	0.8009	0.7810	0.8725	0.8525
	Validation	0.6667	0.7353	0.6932	0.7010	0.8000	0.7273
	Test	0.6234	0.6857	0.6429	0.6545	0.8136	0.7059
XGBoost	Train	0.8654	0.7857	0.8407	0.8255	0.9000	0.8824
	Validation	0.7407	0.8235	0.7727	0.7821	0.8696	0.8000
	Test	0.7013	0.7429	0.7143	0.7221	0.8571	0.7714



(a) Randomly selected six reports (Level 1: multiple-vehicles, level 0: single vehicle)

saturday january approximately hrs dty prejean single crash brule guillot intersection talbot ave upon arrival dty prejean observed white honda civic bearing license plate zed dty prejean spoke james leboeuf jr traveling eastbound brule guillot bad curve intersection brule guillot talbot ave dty prejean observed lost control due speed curve caused back hydroplane side time entered ditch area sideways coming rest dty prejean observed stuck ditch also look sustained damage crash james leboeuf jr wanted schriever auto crash location retrieve ditch dty prejean issued james leboeuf jr traffic ticket rs careless operation motor james leboeuf jr released crash location removed ditch dty prejean resumed patrol nothing reporting beau prejean

Label predicted: 0 (98.66%)

Explainer fit: 0.69

dty toby chamberlain area reference single crash upon arrival came contact turned headed shortly making turn began hydroplane went ditch traveling bound approximately mi exited entering ditch bound shoulder came rest result crash damage side rear portion front side portion none occupants involved crash sustained injuries upon arrival daigles towing passing noticed ditch remained scene remove

Label predicted: 0 (98.96%)

Explainer fit: 0.59

approximately hours traveling bound harvey way side entering ditch struck tree front passenger side causing flip side sliding hood approximately feet striking another tree coming rest drove ditch approximately feet striking first tree front passenger tire defective causing steer towards ditch tried correct also applied brakes locked continued head towards ditch diver even though roads wet raining hydroplane officers opinion traveling approximate speed mph posted mph zone due distance slid hood roof striking first tree passenger require medical treatment time accident required medical treatment refused transport emergency medical service sustained significant cut abrasions upper part arm appears injuries crawling side window sue upside removed owner

Label predicted: 0 (98.74%)

Explainer fit: 0.89

traveling block verot school attempted come stop due front stopped waiting make turn began hydroplane applied brakes stop enough time rear ended complete stop southbound block verot school waiting make turn onto tolson rear ended usps mail truck injuries reported witnesses located completed written issued citation careless operation year located paperwork inside vin number registered valid provided documents state us postal service independent establishment government purchase commercial insurance authorized pay claims liability claims postal service governed federal tort claims act usc also states postal carriers required license plates identification number additional insurance information contact loretta credo district claims coordinator

Label predicted: 1 (53.47%)

Explainer fit: 0.91

(b) Text explanation of Case 3 to Case 6 in the left-hand side (Level 1: multiple-vehicles, level 0: single vehicle)

Fig. 5. Interpretation by keywords from randomly selected crash reports from the test dataset.

(with explanation fit of 0.90) for this case to be considered as single-vehicle crash. The words in the crash narrative identify the class as single vehicle are 'approximately,' 'feet,' 'interstate,' 'began,' and 'hydroplane.' The plots show that the word 'rear' conflicts with single-vehicle crashes, which is obvious as 'rear-end' crashes are usually multiple-vehicle crashes. Case 6 shows that the probability is 0.53 for this case to be considered as multiple-vehicle. High associations are seen for the words 'stop,' 'stopped,' 'brakes,' 'attempted,' and rear. The word 'traveling' shows a negative association. The other cases can also be explained based on the associated values of each word and the overall probability and explanation fit value, a value which represents how much the explanation can be done by using the top 10 words.

Fig. 5b is another display of the mechanisms of the IML algorithm. It shows the condensed form of the crash reports and red and blue identifiers for the classification task for first three cases in the left-hand side of Fig. 5.

6. Conclusions

Studies show that hydroplaning crash outcomes widely vary in single and multiple-vehicle crashes. This study developed a framework of three ML methods to determine the best fit model in identifying vehicle involvement in hydroplaning crashes. The crash characteristics that were identified in this study can help researchers better understand the contributing factors of these crashes. This study assessed the predictive performance of text mining to detect single and multiple-vehicle crashes from hydroplaning crash-related studies. This study also evaluated the efficacy of text mining using free text from crash narratives. The data used in this study included the crash narratives in electronic form for all hydroplaning crashes in Louisiana from 2010 to 2016.

This study achieved two major objectives: 1) developed a framework for applying ML models to unstructured textual content to classify the number of vehicle involvements and 2) applied an interpretable ML framework which can identify the effectiveness of keywords in determining the classification mechanism. The study also included a comprehensive literature review of previous studies that have employed text mining to a traffic crash and occupational injury data. The framework can also be used to conduct other crash-related classifications (e.g., collision type) from the crash narratives. The XGBoost classifiers demonstrated a high prediction power with an accuracy of 71% (test data). This proves that underlying trends or precursors can be revealed and analyzed through ML algorithms. Furthermore, this indicates that quantitative modeling techniques can be used to address safety concerns.

The study has several limitations. Ideally, the classification accuracies should be higher. However, the current scope is limited to examining the performance of the crash narratives in identifying crash types. Additional variables such as vehicle and crash characteristics can improve the performance of the current model. Furthermore, the findings from this study are dependent on the accuracy of the information provided in the text of police reports. Future research should aim to develop a 'traffic crash lexicon' with words that are highly associated with crash characteristics. The development of such lexicon will also improve the model performance.

CRediT authorship contribution statement

Subasish Das: Conceptualization, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Anandi Dutta:** Writing - original draft, Writing - review & editing. **Kakan Dey:** Writing - original draft, Writing - review & editing. **Mohammad Jalayer:** Writing - original draft, Writing - review & editing. **Abhisek Mudgal:** Writing - original draft, Writing - review & editing.

References

Abdat, F., Leclercq, S., Cuny, X., Tissot, C., 2014. Extracting recurrent scenarios from narrative texts using a Bayesian network: application to serious occupational accidents with movement disturbance. Accid. Anal. Prev. 70, 155–166.

Aycock, E., 2008. Hydroplaning: the effect of water on the roadway. collision. The International Compendium for Crash Research 3 (2), 78–85.

Bertke, S.J., Meyers, A.R., Wurzelbacher, S.J., Measure, A., Lampl, M.P., Robins, D., 2016. Comparison of methods for auto-coding causation of injury narratives. Accid. Anal. Prev. 88, 117–123.

Bishop, C., 2011. Pattern Recognition and Machine Learning. Springer.

Black, G.W., Jackson, L.E., Institute of Transportation Engineers (ITE), 2000. Pavement surface water phenomena and traffic safety. ITE Journal 70 (2), 32–37.

Bondy, J., Lipscomb, H., Guarini, K., Glazner, J.E., 2005. Methods for using narrative text from injury reports to identify factors contributing to construction injury. Am. J. Ind. Med. 48, 373–380.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brown, D.E., 2016. Text mining the contributors to rail accidents. IEEE Trans. Intell. Transp. Syst. 17, 346–355.

Bunn, T.L., Slavova, S., Hall, L., 2008. Narrative text analysis of Kentucky tractor fatality reports. Accid. Anal. Prev. 40, 419–425.

Chatterjee, S., 1998. A Connectionist Approach for Classifying Accident Narratives. pp. 1–345 (Theses and Dissertations Available from ProQuest).

Chen, W., Wheeler, K.K., Lin, S., Huang, Y., Xiang, H., 2016. Computerized "learn-as-you-go" classification of traumatic brain injuries using NEISS narrative data. Accid. Anal. Prev. 89, 111–117.

Cornejo-Bueno, L., Borge, J.N., Alexandre, E., Hessner, K., Salcedo-Sanz, S., 2016. Accurate estimation of significant wave height with Support Vector Regression algorithms and marine radar images. Coast. Eng. 114, 233–243.

Das, S., Mudgal, A., Dutta, A., Geedipally, S.R., 2018. Vehicle consumer complaint reports involving severe incidents: mining large contingency tables. Transportation Research Record: Journal of the Transportation Research Board 2672 (32), 72–82.

Enustun, N., 1976. Study to Identify Hydroplaning Accidents (p. 48 p).

Federal Highway Administration, d. Safety and pavement friction. https://safety.fhwa.dot.gov/roadway_dept/pavement_friction/. (Accessed 9 June 2019).

Fisher, A., Rudin, C., Dominici, F., 2018. Model class reliance: variable importance measures for any machine learning model class, from the 'Rashomon' perspective. http://arxiv.o

History of SVM. http://www.svms.org/history.html. (Accessed July 2016).

Traction friction of tires by Ron Kurtus - physics lessons: school for champions. https://www.school-for-champions.com/science/friction_rolling_traction.htm#.XSTX4uhKiUk. (Accessed 9 July 2019).

Jayasooriya, W., Gunaratne, M., 2014. Evaluation of widely used hydroplaning risk prediction methods using Florida's PAST CRASH DATA. Trans. Res. Rec.: J. Transp. Res. Board (2457), 140–150.

Jin, W., Srihari, R.K., Hay Ho, H., Wu, X., 2007. Improving Knowledge Discovery in Document Collections Through Combining Text Retrieval and Link Analysis Techniques.

Khattak, A.J., Kantor, P., F. M. Council, 1998. Role of adverse weather in key crash types on limited-access: roadways implications for advanced weather systems. Transp. Res. Rec. 1621 (1), 10–19.

Labatut, V., Cherifi, H., 2011. Accuracy measures for the comparison of classifiers. the Proceedings of the fifth International Conference on Information Technology, p. 3790.

Lundberg, S., Lee, S., 2016. An unexpected unity among methods for interpreting model predictions. http://arxiv.org/abs/1611.07478. (Accessed 27 July 2018).

Marucci-Wellman, H., Lehto, M., Corns, H., 2011. A combined fuzzy and naïve Bayesian strategy can be used to assign event codes to injury narratives. Inj. Prev. 17, 407–414.

Marucci-Wellman, H.R., Lehto, M.R., Corns, H.L., 2015. A practical tool for public health surveillance: semi-automated coding of short injury narratives from large administrative databases using naïve Bayes algorithms. Accid. Anal. Prev. 84, 165–176.

Molnar, C., 2018. Interpretable machine learning: a guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/. (Accessed 27 July 2018).

Nayak, R., Piyatrapoomi, N., Weligamage, J., 2010. Application of text mining in analysing road crashes for road asset management. Eng. Asset Lifecycle Manag. 49–58.

Pedersen, T., Benesty, M., d. LIME: Local Interpretable Model-Agnostic Explanations. R package version 0.4.0. https://CRAN.R-project.org/package=lime. (Accessed 27 July 2018).

Pollack, K.M., Yee, N., Canham-Chervak, M., Rossen, L., Bachynski, K.E., Baker, S.P., 2013. Narrative text analysis to identify technologies to prevent motor vehicle crashes: examples from military vehicles. J. Saf. Res. 44, 45–49.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you? Explaining the predictions of any classifier. The Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM.

Smith, G.S., Timmons, R.A., Lombardi, D.A., Mamidi, D.K., Matz, S., Courtney, T.K., Perry, M. J., 2006. Work-related ladder fall fractures: identification and diagnosis validation using narrative text. Accid. Anal. Prev. 38, 973–980.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222.

Sorock, G.S., Ranney, T.A., Lehto, M.R., 1996. Motor vehicle crashes in roadway construction workzones: an analysis using narrative text from insurance claims. Accid. Anal. Prev. 28, 131–138.

Vallmuur, K., Marucci-Wellman, H.R., Taylor, J.A., Lehto, M., Corns, H.L., Smith, G.S., 2016. Harnessing information from injury narratives in the 'big data' era: understanding and applying machine learning for injury surveillance. Inj. Prev. 22, 34–42.

Wang, Z., Shah, A.D., Tate, A.R., Denaxas, S., Shawe-Taylor, J., Hemingway, H., 2012. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One 7 (1), 1–9.

Williams, T.P., Betak, J.F., 2016. Identifying themes in railroad equipment accidents using text mining and text visualization. International Conference on Transportation and Development, pp. 531–537 https://doi.org/10.1061/9780784479926.049.

Williamson, A., Feyer, A.M., Stout, N., Driscoll, T., Usher, H., 2001. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. Inj. Prev. 7, 15–20.

Yeager, R.W., 1974. Tire hydroplaning: testing, analysis, and design. In: Hays, D.F., Browne, A.L. (Eds.), The Physics of Tire Traction: Theory and Experiment. Springer US, Boston, MA, pp. 25–63.

Zhou, F., Hu, S., Chrysler, S.T., Kim, Y., Damnjanovic, I., Talebpour, A., Espejo, A., 2019. Optimization of lateral wandering of automated vehicles to reduce hydroplaning potential and to improve pavement life. Transp. Res. Rec., 0361198119853560